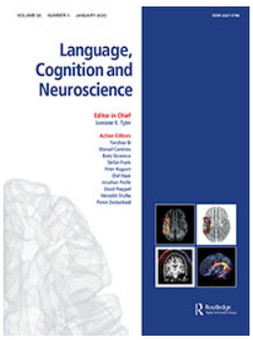
How speakers adapt object descriptions to listeners under load

Vogels, Jorrig; Howcroft, David M.; Tourtouri, Elli; Demberg, Vera

Published in:
Language, Cognition and Neuroscience

DOI:
10.1080/23273798.2019.1648839

Publication date:
2020



Routledge
Taylor & Francis Group

Language, Cognition and Neuroscience

ISSN: 2327-3798 (Print) 2327-3801 (Online) Journal homepage: https://www.tandfonline.com/loi/plcp21

How speakers adapt object descriptions to listeners under load

Jorrig Vogels, David M. Howcroft, Elli Tourtouri & Vera Demberg

To cite this article: Jorrig Vogels, David M. Howcroft, Elli Tourtouri & Vera Demberg (2020) How speakers adapt object descriptions to listeners under load, Language, Cognition and Neuroscience, 35:1, 78-92, DOI: 10.1080/23273798.2019.1648839

To link to this article: https://doi.org/10.1080/23273798.2019.1648839

Published online: 01 Aug 2019.

LANGUAGE, COGNITION AND NEUROSCIENCE
2020, VOL. 35, NO. 1, 78–92
https://doi.org/10.1080/23273798.2019.1648839

Routledge
Taylor & Francis Group

REGULAR ARTICLE

OPEN ACCESS Check for updates

How speakers adapt object descriptions to listeners under load

Jorrig Vogels [a], David M. Howcroft [b], Elli Tourtouri [b] and Vera Demberg [b,c]

[a]Faculty of Arts, Semantics and Cognition — Neurolinguistics and Language Development, Center for Language and Cognition, University of Groningen, Groningen, The Netherlands; [b]Department of Language Science and Technology, Saarland University, Saarbrücken, Germany; [c]Department of Mathematics and Computer Science, Saarland University, Saarbrücken, Germany

ABSTRACT

A controversial issue in psycholinguistics is the degree to which speakers employ *audience design* during language production. Hypothesising that a consideration of the listener's needs is particularly relevant when the listener is under cognitive load, we had speakers describe objects for a listener performing an easy or a difficult simulated driving task. We predicted that speakers would introduce more redundancy in their descriptions in the difficult driving task, thereby accommodating the listener's reduced cognitive capacity. The results showed that speakers did not adapt their descriptions to a change in the listener's cognitive load. However, speakers who had experienced the driving task themselves before and who were presented with the difficult driving task first were more redundant than other speakers. These findings may suggest that speakers only consider the listener's needs in the presence of strong enough cues, and do not update their beliefs about these needs during the task.

ARTICLE HISTORY
Received 6 June 2018
Accepted 16 July 2019

KEYWORDS
Audience design; cognitive load; information density; language production; overspecification; simulated driving

Introduction

For successful communication, it is important that speaker and listener have established a common ground (Clark, 1996; Clark & Wilkes-Gibbs, 1986). For example, a speaker saying "please give me the green chair" needs to have made sure, among other things, that there is an object near the listener that can be uniquely identified by the referring expression "the green chair". If the listener sees only one chair, mentioning "green" is redundant; if the listener sees more than one green chair, the expression may be underspecified. An important question is whether and how speakers adapt referring expressions to their listener's perspective. This type of adaptation is generally called *audience design*, and there has been a long debate about the degree to which speakers employ audience design when making linguistic choices. It is generally accepted that speakers adapt their language to their addressee's perspective at least at a crude level (e.g. Galati & Brennan, 2010), but it is less clear which cues trigger speakers to explicitly consider the listener's needs.

This paper is concerned with the case where a speaker must uniquely identify a referent for a listener who is experiencing an increased cognitive load. The processing capacity of addressees is likely to vary inversely with their extralinguistic cognitive load. For instance, when an addressee is performing a secondary task while receiving a message, this task will reduce their processing capacity, leaving less capacity for processing the information in the message. If speakers are sensitive to the cognitive load of their listeners, they should adapt their language use to remain comprehensible. If speakers do not take into account their listeners' cognitive state, no adaptation is expected when the listener is under increased cognitive load. There are different views in the psycholinguistic literature on the degree to which speakers take their addressee's needs into account in their referential choices, to which we will now turn.

Addressee-oriented and egocentric views of reference production

According to the addressee-oriented view on reference production (e.g. Brennan & Clark, 1996; Clark, 1996; Clark & Wilkes-Gibbs, 1986), cooperative speakers include as much information as needed for the listener to pick out the correct referent. The idea that speakers take into account the informativeness of their utterances for their addressee is expressed by Grice's (1975) Maxim of Quantity: (1) Make your contribution as informative as is required, but (2) do not make your contribution more informative than is required. Such inferences made by

CONTACT Jorrig Vogels j.vogels@rug.nl

the speaker about how informative an utterance will be for the listener are incorporated in some theories of referring expression production, which assume that in selecting a referring expression speakers take into account how the listener will interpret it (e.g. Gundel, Hedberg, & Zacharski, 1993; Hendriks, 2016), as well as in information-theoretic models of pragmatic reasoning such as the Rational Speech Act model (Frank & Goodman, 2012; Goodman & Frank, 2016). Within the psycholinguistic literature on reference, there is evidence that speakers are at least moderately Gricean. For example, speakers reduce referring expressions when they refer repeatedly to the same object while speaking to the same addressee (Brennan & Clark, 1996; Clark & Wilkes-Gibbs, 1986; Horton & Gerrig, 2005), but can adapt relatively quickly to a new addressee who is not yet familiar with the object (Galati & Brennan, 2010; Gann & Barr, 2014). In addition, speakers have been found to shorten the duration of parts of their description in response to characteristics of the listener (e.g. Arnold, Kahn, & Pancani, 2012; Rosa, Finch, Bergeson, & Arnold, 2015).

However, other research has emphasised that speakers are not always Gricean (e.g. Bard et al., 2000; Dell & Brown, 1991; Pickering & Garrod, 2004). For example, speakers often refer to information that is not available to their addressee (Horton & Keysar, 1996; Wardlow Lane, Groisman, & Ferreira, 2006), and they overspecify their references in cases where this is not considered beneficial for the listener (Engelhardt, Bailey, & Ferreira, 2006; Koolen, Gatt, Goudbeek, & Krahmer, 2011). It has been suggested that maintaining a detailed mental model of the addressee's needs is cognitively costly (Bard et al., 2000; Dell & Brown, 1991; Goudbeek & Krahmer, 2011; Horton & Keysar, 1996; Rossnagel, 2000), and generally not very efficient, given that the speaker's and the listener's knowledge are often closely aligned (Brennan & Hanna, 2009; Galati & Brennan, 2010). Hence, speakers could use their own knowledge as a proxy for their addressee's (Pickering & Garrod, 2004). There is also empirical evidence that speakers base referential choices mainly on their own model of the discourse rather than on an explicit consideration of the listener's perspective (e.g. Bard & Aylett, 2005; Fukumura & van Gompel, 2012; Vogels, Krahmer, & Maes, 2015). Thus, despite the findings suggesting speaker adaptation in reference, reference production may still be at least partly insensitive to the listener's perspective; speakers seem to often produce referring expressions based on their own, egocentric preferences.

In sum, it is generally accepted that speakers can and do adapt referring expressions to their addressees' needs, but which cues are instrumental in triggering such adaptive behaviour is still poorly understood. The question, then, is which situations trigger perspective taking in reference production, and which do not. One possibility is that speakers only take into account the addressee's knowledge in their referring expressions when there is a clear risk of misinterpretation or when it is very important that the message be understood correctly (e.g. Arts, Maes, Noordman, & Jansen, 2011; Watson, Arnold, & Tanenhaus, 2008; but cf. Jucks, Becker, & Bromme, 2008). For instance, Arts and colleagues found that speakers produced more overspecified descriptions of geometrical objects when they were told that they were engaged in a long-distance surgery than when no cover story was given. In addition, speakers have been found to produce longer descriptions when addressees are distracted (Rosa et al., 2015; but cf. Kuhlen & Brennan, 2010). Therefore, speakers may employ audience design whenever there are strong enough cues that adaptation is necessary to ensure successful communication (e.g. Fukumura & van Gompel, 2012; Pickering & Garrod, 2004).

However, one problem in investigating audience design is that it is difficult to know when the use of a certain linguistic form constitutes evidence for taking into account the listener's perspective and when it does not. For example, according to the Maxim of Quantity producing an overspecified description is not cooperative, since it gives more information than necessary, and hence overspecifications have been considered evidence of egocentricity (Engelhardt et al., 2006). On the other hand, an increase in the number of overspecifications has also been taken to reflect audience design where it occurred after a change in the communicative situation (see the Arts et al., 2011 study on referring expressions in a long-distance surgery cited above).

Uniform information density

To provide us with a clear prediction of what listener adaptation in referential descriptions should look like linguistically, we turn to the Uniform Information Density hypothesis (UID; e.g. Levy & Jaeger, 2007). According to UID, speakers strive to distribute information equally across their utterances. To this end, they are predicted to make linguistic choices (where more than one alternative is permitted by the grammar of the language) that maximise the amount of information conveyed to a listener within a threshold of comprehensibility, also known as the channel capacity. In Information Theory, the channel capacity is defined as the rate at which information can be transmitted successfully (Shannon, 1948). In principle, channel capacity depends on every aspect of the channel, from its conception in the speaker's mind to

its interpretation by the listener (Jaeger, 2010). For our purposes, we consider channel capacity at the receiver's end, and take it to be the amount of cognitive resources available to the listener for processing linguistic material. In general, speakers like to be parsimonious (the *Principle of least effort*; Zipf, 1949), and hence strive to transmit as much information as possible with the least possible effort. However, a rate of information transmission exceeding the channel capacity may result in information loss. It follows that information is distributed optimally for both speaker and listener when its rate of transmission is uniformly close to the channel capacity, but not exceeding it.

There is an increasing amount of evidence that UID is an important force driving linguistic choices in language production (e.g. Jaeger, 2010; Mahowald, Fedorenko, Piantadosi, & Gibson, 2013; Piantadosi, Tily, & Gibson, 2012), but it is not yet clear what underlies this behaviour in speakers. Although UID itself is agnostic about the degree to which speakers' linguistic choices are motivated by audience design, one intuitive possibility is that speakers try to minimise listeners' processing difficulty. Peaks in information density may cause interpretation problems (e.g. Van Berkum, 2008; Demberg & Keller, 2008; Federmeier, McLennan, De Ochoa, & Kutas, 2002; Jaeger & Tily, 2011; Levy, 2008; Kutas, DeLong, & Smith, 2011; Smith & Levy, 2013; Xiang & Kuperberg, 2015), and keeping information density uniform close to the channel capacity ensures that listeners will not be overloaded by information. For reference production, UID predicts that parts of a referential description that have a high information density will be spread out over more time, whereas material that has a relatively low information density will be reduced. Indeed, research has shown that where speakers overspecify, the redundant information is shorter in duration than the same information in a minimally specified reference (Engelhardt & Ferreira, 2014). Moreover, redundant information that results in a more uniform reduction of referential entropy (i.e. how many possible referents there are at a certain point in a description) seems to reduce the listener's processing effort (Tourtouri, Delogu, & Crocker, 2017).

Crucially, if speakers are sensitive to the processing capacity of their addressees, they should also adjust the overall information density of their utterances to a level that they expect the addressee to be able to process. More specifically, they should introduce more redundancy in their referring expressions when the listener is experiencing an increased cognitive load. Hence, we hypothesise that, when the addressee is involved in a difficult task that is noticeably reducing their cognitive capacity, speakers will produce more overspecified referential descriptions, thereby distributing distinguishing (and therefore more informative) content over more linguistic units, and hence reducing the overall information density of their utterances. This may help the addressee in selecting the correct referent in the difficult task, because he will have more time and more linguistic cues to identify the referent.

In contrast, an alternative strategy that speakers may employ to aid their addressee is to minimise the time that they disturb the listener, using as short descriptions as possible. This would mean that the information density of their descriptions would actually increase when the listener is under load. Finally, it could be the case that speakers do not adapt to the needs of listeners under load at all, or perhaps only when they have a simple but compelling cue that adaptation is necessary, for example when they have experienced the same cognitive load themselves.

The current study

To investigate these hypotheses, we conducted an experiment in which pairs of participants performed a referential communication task. One participant, the speaker, described objects for the listener, who was performing a secondary driving task in a driving simulator. Driving is a complex cognitive task, and it has been shown that listening or talking and driving at the same time can seriously impact driving performance (Demberg, Sayeed, Mahr, & Müller, 2013; Drews, Pasupathi, & Strayer, 2008). Conversely, there is also some evidence that driving has an impact on language comprehension and production (Becic et al., 2010; Engonopoulos, Sayeed, & Demberg, 2013). In our dual task setting, the listener (henceforth called the "listener-driver") was behind the wheel in a driving simulator, while the speaker (henceforth called the "speaker-passenger") was in the passenger seat and described an object amidst several other objects, appearing above the road on the simulator screen. The listener-driver's task was to identify the object that was referred to. Extra-linguistic cognitive load of the listener was manipulated by setting the driving task to either easy or difficult. In the easy condition, the listener-driver had a perfectly straight road in front of him and had to do practically no steering. In the difficult condition, the listener-driver had to use the steering wheel to keep two vertical bars appearing on the road perfectly overlapping (the ConTRe task; Mahr, Feld, Moniri, & Math, 2012). As the ConTRe task has been shown to increase cognitive load (Demberg et al., 2013), the listener-driver's processing capacity for incoming linguistic material will be decreased during this task. The question is whether

speakers are sensitive to this decrease in processing capacity in performing the referential communication task. Because speakers may only take into account their listener's cognitive state when they have concrete evidence that it is different from their own, and hence cannot use a model of their own cognitive state as a proxy, we had the speaker-passenger and the listener-driver switch roles halfway through the experiment, allowing us to investigate the influence of a speaker's first-hand experience with the cognitively taxing task on their later descriptions for someone performing that same task.

To investigate how speakers accommodate their listener's increased cognitive load in their referential descriptions, we measured the degree of referential overspecification in these descriptions. Each object to be referred to could be uniquely identified by mentioning either one or two properties (i.e. a minimal description), and expressions that used more properties than that (i.e. mentioned redundant attributes) were considered overspecified (e.g. Koolen, Goudbeek, & Krahmer, 2013). Redundant attributes were never fully distinguishing. In addition, we analysed the general information density (amount of information per linguistic unit) of the descriptions, both as the number of words per attribute and as the average word duration. We also analysed the speech rate with which redundant and non-redundant modifiers were produced within overspecified descriptions. The rationale behind this was that, given the prediction of UID that linguistic material carrying less information will be reduced, redundant attributes in the descriptions are likely to have a specifically reduced pronunciation (Engelhardt & Ferreira, 2014). The rate of reduction may then be diminished when the addressee is under load. Furthermore, we analysed the speaker-passenger's speech onset latency to determine whether speakers also take longer to plan their description when they show signs of adaptation. Finally, we examined the listener-driver's response accuracy and driving accuracy, to see how their performance was influenced by different types of descriptions.

We predicted that when listeners are under an increased cognitive load, speakers decrease the information density of their utterances. Generally, we predicted descriptions to become less informationally dense in the difficult driving task than in the easy driving task, i.e. the same referential information is spread out over more time or more words. More specifically, we expected speaker-passengers to be more redundant in their descriptions, and include more modifiers than necessary in the difficult than in the easy driving task to give the listener-driver more cues and more time to process the descriptions. Alternatively, if speakers shorten their descriptions so as to cause as little disturbance to the driving task as possible, we would instead expect an increase in the information density of the descriptions. If such changes in either redundancy or description length are the result of an effortful process of audience design, we would also expect speakers to take longer to plan their descriptions, resulting in a longer speech onset latency. In addition, if adaptation strategies are beneficial for listeners under load, we expect performance on the referential task and/or on the driving task to be better when descriptions are adapted to the listener. Finally, if speakers take into account their listener's cognitive state only when they have concrete evidence that it is different from their own, we might find evidence for adaptation primarily in speaker-passengers who did the driving task first compared to speaker-passengers who did not yet experience the driving task themselves.

Method

Participants

Twenty-five pairs of Saarland University students, with mean age 23.4 (*SD* 3.9), participated in our experiment and were paid €10. Twenty-nine participants were women and the rest were men. Two pairs did not switch roles halfway through the experiment, resulting in referring expression data for 48 participants in total. All participants provided written consent and their data were fully anonymised.

Materials

The stimulus materials were based on those used in Koolen et al. (2011).[1] They consisted of scenes containing 7 images in a 2×4 grid, where each grid position was numbered 1–8 (i.e. one grid position remained empty). The target image was identified for the speaker by number, with this number appearing on a separate display not visible to the driver. The images in a scene were furniture, taken from the Object Databank produced by Michael Tarr's lab.[2] The image set in this domain is highly systematic, consisting of four different object types (chair, sofa, desk, fan) in four different colours (blue, red, green, grey), four different orientations (front-, back-, left-, right-facing), and two different sizes (large, small). For the present experiment, we left out the backward facing objects because we judged them to be less clear in our particular setup, thus leaving us with three different orientations.

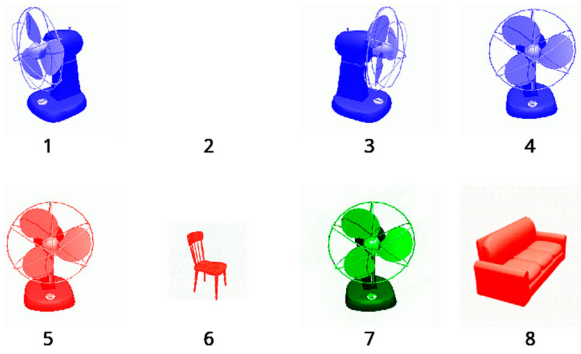


Figure 1. Higher resolution version of an example stimulus with minimal description length 2 to describe image #4.

We created 88 scenes, constructed in such a way that either one ($n = 36$) or two ($n = 52$) modifiers were minimally required for the listener to pick out the target image. For example, in Figure 1, image #4 is identifiable by the attributes "orientation = front" and "colour = blue", allowing for the minimal description *der blaue Ventilator, der nach vorne zeigt* ("the blue fan facing front"). In addition, we created 32 filler items, which consisted of 8 items for which no attributes were required for a minimal description, and 24 for which the mention of all three attributes was needed, resulting in a total of 120 scenes.

Since every participant acted once as the speaker-passenger and once as the listener-driver in the experiment, we created two item lists to ensure that participants did not see the same item twice. Each list consisted of 60 trials. The target images were selected such that all combinations of the minimally required attributes occurred equally often on a list, and that the same image appeared only once as the target referent. In addition, we created 8 practice trials (the same on the two lists), which were similar but not identical to the experimental items.

Driving simulator

The experiment was run in a driving simulator consisting of two front seats, dashboard, steering wheel and gas and brake pedals taken from a real car. We used the OpenDS 3.0 software (https://www.opends.dfki.de; Math, Mahr, Moniri, & Müller, 2012) to provide the driving environment, which was projected on three large panels positioned in an approximately 180° curve around the car (see Figure 2). The driver's seat was aligned with the centre of the middle panel. The presentation of the stimuli was controlled from a separate PC using the Experiment Builder software (https://www.sr-research.com), which communicated with the driving simulator software over a serial port connection. The speaker-passenger received instructions regarding the identity of the referent to be described via a second display (an iPad) connected to the Experiment Builder PC using duet display (https://www.duetdisplay.com/). A microphone was mounted on the dashboard in front of the passenger, and an Eyelink 1000 Plus eye tracker was placed just behind the steering wheel. The eye tracker was used to record the listener-driver's eye movements and pupil size as on-line measures of the linguistic processing of the image descriptions. However, as the current paper is concerned with linguistic choices in production, we will not report on these measures here.

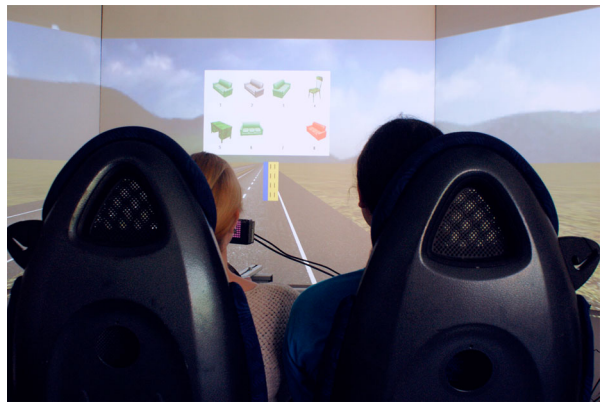


Figure 2. The driving simulator used for this experiment. Speaker-passengers described a target image from the image array displayed above the road. Listener-drivers had to follow the yellow bar back and forth across the road while identifying the intended referent by number.

Procedure

Two participants were randomly assigned to the roles of speaker-passenger and listener-driver using a coin toss. The listener-driver sat in the driver's position in the driving simulator. The speaker-passenger then sat down on the passenger seat and received an iPad. In each trial the iPad displayed a number identifying a target image to the speaker. The speaker-passenger's task was to describe the target image in such a way that the listener-driver could identify the correct image from among the distractor images. Speaker-passengers were free to describe the images in any way they wanted, except that they were told that they were not allowed to use pointing or mention the target's location in the scene or its assigned number. Every ten trials, a question about the perceived cognitive load of the listener-driver (*Wie abgelenkt finden Sie den Fahrer jetzt?* "How distracted do you find the driver at the moment?") appeared on the iPad for the speaker-passenger to answer. This question was only intended to keep the speaker-passenger aware of the cognitive state of the listener-driver, and answers were not recorded.

The listener-driver's task was to listen to the speaker-passenger's descriptions, and identify the correct referents in the scenes that appeared above the road in the driving simulator. If the description was not clear, the listener-driver was allowed to ask for clarification. After each trial, the listener-driver had to say aloud the number of the referent based on the description provided. The experimenter then recorded this number. Each trial had a time limit of 15 s, after which the next trial was automatically started.

Besides the referential communication task, the listener-driver had to perform a secondary driving task simultaneously. There were two driving conditions: In the easy driving condition, listener-drivers did not have to do anything else but keeping the car straight on a straight road. In the difficult driving condition, listener-drivers had to perform the ConTRe task while identifying the referents. In this task, a blue and a yellow vertical bar are present on the road directly in front of the driver, at a fixed distance. The yellow bar moves randomly back and forth across the road. The blue bar is controlled by the steering wheel and is always in the middle of the driver's view. The driver's task is to keep turning the steering wheel (and thereby the car) in such a way that the (narrower) blue bar is centred in the yellow bar as much as possible (see Figure 2).

The experiment started with a short training block, in which the different tasks (referential communication, difficult driving, simultaneous driving and communicating) were practised separately. After the training session, the eye tracker was calibrated. The actual experiment consisted of two blocks of 30 trials, each assigned to one of the driving conditions. Half of the participants got the easy driving condition in the first block and the difficult driving condition in the second. The order was reversed for the other half of the participants. After the two blocks were completed, the speaker-passenger and listener-driver switched roles, and the experiment was repeated (but with a different item list). For all participant pairs but one, the new listener-driver got the same order of conditions as the listener-driver before the role switch. Given that we had 25 pairs, however, there was one pair in which the first listener-driver got the easy driving condition first, and the second listener-driver got the difficult driving condition first. A schematic representation of the design is presented in Figure 3.

All speech, both that produced by the speaker-passenger and that produced by the driver-listener, was recorded using a single, dash-mounted microphone. Each complete experiment, in which both participants had been speaker-passenger and listener-driver once, lasted approximately 1.5 h.

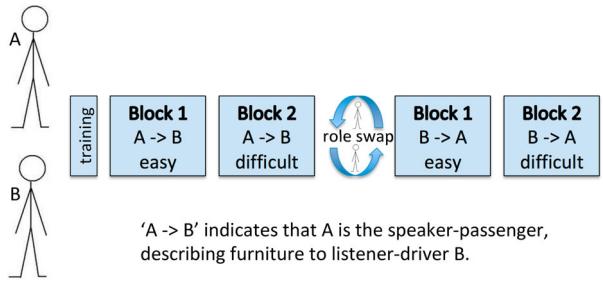


Figure 3. Schematic representation of the experimental design. The order of the blocks was counterbalanced across participants.

Design and coding

Varying driving condition (easy driving, difficult driving) as a within-participants factor and role order (speak-first, drive-first) as a between-participants factor resulted in a 2×2 mixed design. Block (1, 2) and the number of minimally required attributes (1, 2) were included as control factors. The main dependent variable was the degree of overspecification in referential expressions. Other linguistic measures analysed for the speaker-passenger's utterances were number of words per description (relative to the number of attributes mentioned), average word duration, modifier speech rate within overspecified descriptions, and speech onset latency. For the listener-driver, we analysed referent identification accuracy, identification speed, and driving accuracy (steering deviation).

Only the first referring expression in each experimental trial was considered. Any later adjustments triggered by listener feedback (55 cases in total) were ignored. We further excluded descriptions that fell in one of the following categories: descriptions with major (listener) interruptions or other disturbances (33 cases); descriptions of the wrong image (12 cases); references to earlier trials or other objects in the scene (11 cases); attempts at humour or too much creativity (10 cases); experiment errors (5 cases); descriptions that were not completed within the time limit (3 cases); and mentioning the number of the image (1 case). Finally, any descriptions using the right number of attributes but the wrong attribute type (e.g. mentioning colour and size where colour and orientation are the distinguishing properties) were excluded (36 cases). These steps resulted in a data loss of 4.9%, yielding 2008 referring expressions for analysis.

Analysis and results

Degree of overspecification

The degree of overspecification was determined by counting the number of redundant attributes in each

referring expression. Any mentioned property that was not necessary to uniquely identify the referent was considered redundant (cf. Koolen et al., 2013). For any attribute required for a minimal description that was not mentioned (i.e. underspecification), we subtracted 1 from this number. We ran a cumulative link mixed effects model on the degree of overspecification, using the ordinal package in R (Christensen, 2015). Cumulative link models allow for the analysis of ordinal dependent variables. The initial model included driving condition (easy driving/difficult driving), role order (speak first/drive first), and their interaction as main predictors, and block (1/2) and minimally required attributes (1/2) as control predictors. All predictors were centred to reduce collinearity. We started by fitting a model including random intercepts for participants and items, and by-participant and by-item random slopes for driving condition.[3] Next, we used model selection to simplify the model, starting with removing the random correlations and removing random components with low variance (see Bates, Kliegl, Vasishth, & Baayen, 2015), and then removing fixed effects of the control predictors. We used a Likelihood Ratio test to determine if each removal was justified. The main predictors and their interaction always remained in the model.

Overall, we found a high rate of overspecification in the referring expressions: 55% of referring expressions were overspecified, i.e. mentioned at least one attribute that was not required to distinguish the target referent from the competitors. This figure is comparable to the rate of overspecification found in the D-TUNA corpus (53.6%), which used the same type of materials (Koolen et al., 2011). Underspecification was very uncommon: Only 0.4% of referring expressions (9 instances) lacked one or more attributes that were necessary to distinguish the referent (1.9% when counting the wrongly specified referring expressions that we removed earlier).

We analysed whether the degree of overspecification (average number of redundant attributes) was different between the two driving conditions, and whether this interacted with role order. As there were very few cases of underspecification, we removed this value from the ordinal response variable. The final model showed a significant main effect of the minimal number of required attributes (a lower degree of overspecification when two attributes were already required), but no significant effects of driving condition and role order, and no interaction between the two (see Table 1). However, inspecting the data in more detail revealed a marked difference between the first and the second block of the experiment. Figure 4 illustrates the complex relationship between block, driving condition and whether the speaker or the listener role was taken first. Block 1 shows a higher degree of overspecification in the difficult driving condition, at least for participants who drove first. In Block 2, this trend seems to be reversed: Participants who had driven first produced a higher degree of overspecification in the easy driving condition than in the difficult driving condition.

Table 1. Cumulative link mixed effect model output for the degree of overspecification. Threshold coefficients are the intercepts for each of the response categories.

	β	*SE*	*z*	*p*
Threshold coefficients				
0|1	−0.4816	0.2286	−2.1068	0.0351
1|2	3.7023	0.255	14.5162	<.001***
Coefficients				
Driving condition	0.0416	0.1512	0.2753	0.7831
Role order	0.6606	0.4524	1.4603	0.1442
Minimal number of required attributes	−2.1248	0.3434	−6.1875	<.001***
Driving condition : Role order	−0.4258	0.3028	−1.4064	0.1596

The pattern in Figure 4 may be explained when we consider that driving difficulty was manipulated blockwise, i.e. whenever Block 1 had the easy driving condition, Block 2 had the difficult driving condition, and vice versa. Thus, speakers may have been copying their referential strategy from Block 1 into Block 2, even though Block 2 had the other driving condition. This becomes visually clear when the bars for the easy and difficult driving conditions are swapped for Block 2, as was done in Figure 4.

To account for this pattern of results, we ran a post-hoc analysis in which we added a predictor condition order, indicating whether the easy or the difficult driving condition was done first. The results in Table 2 show a significant positive interaction between condition order and role order. Running paired comparisons for the interaction revealed that when the speaker had driven first, there was a significant main effect of condition order ($\beta = 1.2216$; $SE = 0.4836$; $p = .01$), confirming that speakers who had experienced the driving task and who were presented with the difficult driving condition as the first driving task overspecified their descriptions to a greater degree than speakers who had experienced the driving task but were presented with the easy driving condition first. There was no effect of condition order for speakers who had not driven first ($\beta = -0.1539$; $SE = 0.3600$; $p = .67$). Because the predictors condition order and role order both varied only between participants and between items, the random effects structure for the model shown in Table 2 only included a by-participant random slope for driving condition and random intercepts for participant and item.

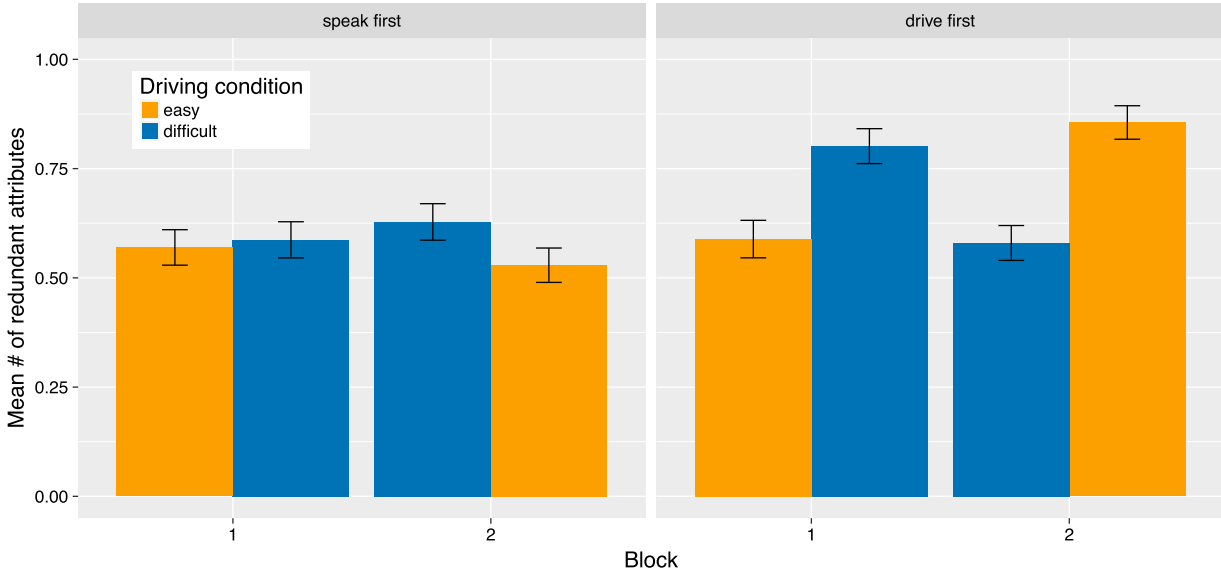


Figure 4. Degree of overspecification in speakers who described first (left) and speakers who had driven first (right) by driving condition and block. For an individual participant pair, if Block 1 had the easy driving condition, Block 2 had the difficult driving condition and vice versa, as indicated by the swapping of the colours in Block 2.

Table 2. Cumulative link mixed effect model output for the degree of overspecification with condition order as main predictor. Threshold coefficients are the intercepts for each of the response categories.

	β	*SE*	*z*	*p*
Threshold coefficients				
0|1	−0.4689	0.2202	−2.1298	0.0332
1|2	3.7144	0.2477	14.9977	<.001***
Coefficients				
Condition order	0.5402	0.3000	1.8008	0.0717
Role order	0.6347	0.4354	1.4579	0.1449
Minimal number of required attributes	−2.1240	0.3432	−6.1887	<.001***
Condition	0.0417	0.1543	0.2704	0.7868
Condition order : Role order	1.3520	0.6000	2.2536	0.0242*

Description length

For description length, we first analysed the number of words in the descriptions, divided by the number of attributes mentioned (including type, and excluding material following listener feedback). Second, we analysed the average word duration in seconds within a description. We used the software package MAUS (http://www.bas.uni-muenchen.de/Bas/BasMAUS.html) to automatically find the word boundaries in the speech signal based on the manual transcription. Start and end of each description were corrected manually. We log transformed both variables to make the data more normally distributed and ran linear mixed effect model analyses using the same procedure as described above. Both analyses showed a small but significant effect of minimally required attributes, with more words per attribute and shorter word durations in descriptions where two attributes were minimally required as compared to descriptions that required the mention of only one attribute (number of words: 2.08 vs. 1.96 words; $\beta = 0.0923$; $SE = 0.0404$; $p = .02$; word duration: 0.42 vs. 0.45 s; $\beta = -0.0534$; $SE = 0.0185$; $p = .005$). However, there were no significant effects of driving condition, role order, or their interaction in either analysis (all ps > .05). Rerunning the analyses with condition order as an additional predictor yielded very similar results.

Modifier speech rate

Since adaptation effects may manifest themselves more locally, we also analysed the rate at which both redundant and non-redundant modifiers were spoken within overspecified descriptions, in syllables per second. The difference between redundant and non-redundant modifiers was not statistically significant, and there was no significant interaction with driving condition (both ps > .1). However, there was a significant interaction between driving condition and role order, such that, irrespective of modifier redundancy, speakers who had driven first pronounced modifiers faster in the difficult than in the easy driving condition (see Table 3). An analysis including condition order as a predictor did not show effects of redundancy ($p = .47$).

Speech onset latency

Next, we analysed the mean latency with which speakers started their description (in seconds). Again, a log transformation was applied, and linear mixed effect models

Table 3. Linear mixed effect model output for log modifier speech rate (overspecified descriptions only).

	β	*SE*	*t*	*p*
(Intercept)	1.4526	0.034	42.7242	<.001***
Modifier redundancy	0.0276	0.0343	0.8064	0.4223
Condition	0.0490	0.0159	3.0769	0.0021**
Role order	−0.0025	0.0675	−0.0374	0.9703
Modifier redundancy : Condition	0.0301	0.0318	0.9466	0.3440
Modifier redundancy : Role order	−0.0107	0.0684	−0.1560	0.8764
Condition : Role order	0.0661	0.0321	2.0601	0.0395*
Modifier redundancy : Condition : Role order	0.0657	0.0641	1.0248	0.3056

were run in the same way as described above. We hypothesised that speech onset latencies would be longer if speakers adjusted their descriptions to their addressee's level of cognitive load. However, since we did not find evidence for trial-by-trial adaptation, we instead investigated the effect of overspecification in general on speech onset latency. If overspecifications are the result of effortful audience design, we expect longer latencies for overspecified descriptions. Alternatively, if overspecification is easier for the speaker, we expect shorter latencies for overspecified descriptions. Thus, we ran an analysis including a predictor for the rate of overspecifiation, as well as a predictor for condition order. The final model, presented in Table 4, shows a significant difference between overspecified and minimally specified descriptions, with shorter speech onset latencies for overspecified than for minimally specified descriptions. This suggests that redundant descriptions were not effortfully selected in order to accommodate the listener.

Driver's response and driving accuracy

Finally, we assessed the driver's accuracy and speed (response time in seconds) in identifying the correct object as well as their driving accuracy (deviation between the steering and the reference bar in metres). Linear mixed effect models were run for driving accuracy and response time, and logit mixed effect models were run for response accuracy. In all cases, we included rate of overspecification in the description as an ordinal predictor. To correct for the longer duration of overspecified descriptions, we measured response time from the time point where the description was fully distinguishing (i.e. referential entropy was zero).

Table 4. Linear mixed effect model output for log speech onset latency.

	β	*SE*	*t*	*p*
(Intercept)	0.8590	0.0366	23.4558	<.001***
Condition order	−0.0132	0.0722	−0.1824	0.8561
Role order	−0.0877	0.0728	−1.2036	0.2349
Overspecified vs. minimally specified	0.0311	0.0138	2.2568	0.0252*
Overspecified by 2 vs. 1 attributes	−0.0027	0.0209	−0.1306	0.8965
Condition	−0.0049	0.0153	−0.3227	0.7486
Condition order : Role order	−0.2705	0.1443	−1.8750	0.0674

The drivers' response accuracy was very high, both in the easy and in the difficult driving condition (98.4% and 97.6% correct, respectively), indicating that the passenger's descriptions were largely successful and that participants were attending to the referential task. There was no significant effect of overspecification on response accuracy (overspecified vs. minimally specified: $p = .76$; 2 vs. 1 redundant modifier: $p = .85$).

Next, we tested the hypothesis that drivers in the difficult driving condition would be relatively faster to identify the correct referent when the description was overspecified than when it was not. This hypothesis was not confirmed: The driver's response time was not significantly different between descriptions with two redundant modifiers, descriptions with one redundant modifier, and minimal descriptions ($ps > .1$).

Finally, there was a significant difference in steering performance between the two driving conditions ($\beta = 0.3548$; $SE = 0.0134$; $p < .001$): Drivers were better at keeping the bars overlapped in the easy driving condition (when the yellow bar did not move; mean steering deviation 0.01 m) than in the difficult driving condition (when it moved randomly across the road; mean steering deviation 0.37 m), confirming that the difficult driving condition was indeed more challenging. However, driving accuracy was not significantly different between overspecified and minimally specified descriptions (overspecified vs. minimally specified: $p = .25$; 2 vs. 1 redundant modifier: $p = .48$).

General discussion

In this study, we investigated whether speakers adapt their referring expressions to the listener when the listener is noticeably under an increased cognitive load. Speakers described carefully controlled objects to listeners in a simulated driving context. The listener had to identify the correct object based on the description, while performing either an easy driving task involving little to no steering or a difficult driving task resulting in extensive steering movements. Based on the Uniform Information Density hypothesis, we predicted that speakers would incorporate more redundant attributes (properties of the object that do not need to be mentioned for a minimal description distinguishing the object from the other objects in the scene) in their descriptions in the difficult driving condition as compared to the easy driving condition. In this way, they

would reduce the overall information density of their utterances, and hence accommodate listeners under load by giving them more cues and more time to process the information. More generally, we predicted that if speakers aid their listeners by spreading out information over more words or over more time, they would also produce longer descriptions in the difficult driving condition.

Our results do not confirm these predictions. In line with earlier studies on overspecification in language production (e.g. Arts et al., 2011; Engelhardt et al., 2006; Koolen et al., 2011), we found that mentioning redundant attributes was very common: More than half of all referring expressions contained at least one redundant attribute. However, speakers apparently did not use this redundancy actively as a way to accommodate a listener under increased cognitive load in our simulated driving task: They did not adjust the degree of overspecification when drivers switched from an easy driving task to a difficult driving task, or from a difficult driving task to an easy driving task. Instead, the referential strategy from the first block of the experiment seemed to be copied over and reinforced in the second block, even though this block had the other driving condition. More generally, we did not find evidence for our hypothesis that descriptions would be longer (on top of the number of attributes mentioned) in the difficult than in the easy driving condition. Speakers did not increase word duration in the difficult as compared to the easy driving condition either.

However, we did find an interesting between-group difference in the rate of overspecification. Speakers who had both experienced the driving task first and were presented with the difficult driving condition as the first driving task included more redundant attributes in their descriptions than other speakers. Although we only found this effect in a post-hoc analysis, we believe that it does make some interesting suggestions about the degree to which speakers adapt their linguistic choices to their addressees. First, the finding that the group of speakers who produced more redundant descriptions had experienced the driving task themselves first may suggest that speakers do not automatically adapt their referring expressions to their addressees, but are only more redundant when they have compelling evidence that doing so would be important. This would be consistent with accounts of language production stating that speakers only take the listener's perspective into account when there are strong cues that adjustment is necessary (e.g. Fukumura & van Gompel, 2012; Pickering & Garrod, 2004).[4] Having first-hand experience of the addressee's task may be one of the cues triggering audience design. Hence, perspective taking in language production may be triggered by quite literally putting yourself in the listener's shoes.

Second, the finding that speakers did not adjust their level of redundancy between the first and the second block of the experiment, even though the second block had the other driving condition, might suggest that descriptions were affected by the driving difficulty at the start of the experiment, but not by a change in driving difficulty halfway through. If this is true, it would corroborate the view that speakers choose a referential strategy that takes into account the needs of the listeners only on a global level and do not continuously update their beliefs about the listener during interaction. Many studies have found evidence for global adaptation. For example, when people repeatedly tell the same story or refer to the same object, utterances become less detailed, but less so when they are addressed to a new person who has not heard the original story or reference (Brennan & Clark, 1996; Galati & Brennan, 2010). Also, people describe objects differently when talking to friends or insiders than to strangers or outsiders (Isaacs & Clark, 1987; Krauss & Fussell, 1991). The necessity of such adaptations is easy to establish for the speaker, as the addressee's needs are simple and clear from the start (Brennan & Hanna, 2009; Galati & Brennan, 2010). In addition, the assessment of the addressee's needs is also dependent on the expectations that the speaker has about the addressee. For example, Kuhlen and Brennan (2010) found that speakers told jokes with more detail to attentive addressees than to distracted addressees, but not when they expected the addressee to be distracted. Similarly, in the current study, when speakers expected a mentally loaded addressee, they may have acted on this expectation irrespective of whether or not the driving task was actually increasing the addressee's cognitive load, which may be why the order in which the driving conditions were presented to the participants mattered for the degree of redundancy in their descriptions.

Although this interpretation of our findings allows us to maintain that speakers use overspecification as a way to accommodate listeners under load, be it only under specific conditions, other task strategies may have been employed as well. For example, our finding that modifiers were spoken faster in the difficult than in the easy driving condition only for speakers who had done the driving task first might point to a different audience design strategy (cf. Arnold et al., 2012; Rosa et al., 2015): Instead of increasing redundancy and distributing the information across a longer time period, some speakers may have employed a strategy in which part of their descriptions are shortened to not disturb the listener's

driving task too much. The fact that speech rate was similar for redundant and non-redundant modifiers may suggest that the degree of redundancy did not play a role in this referential strategy.

An alternative interpretation of our findings could be that differences in the rate of overspecification are due to speakers' own egocentric preferences. As argued by some researchers (e.g. Arnold et al., 2012; Arnold & Griffin, 2007; Rosa et al., 2015), what may seem like listener accommodation may in fact be the result of speaker-internal processes. For example, a difficult dual-task setting for the addressee may also distract the speaker, which may hinder her capacity to produce an appropriate object description, possibly resulting in increased redundancy. Our results for both speech onset latency and driving accuracy are also consistent with an account where speakers' choices in production are at least influenced by speaker-internal processes. Firstly, we found a significantly shorter latency to begin speaking for overspecified as compared to minimally specified descriptions. A potential explanation is that speakers who intended to be more redundant followed a very simple strategy: to just start describing without paying attention to which modifiers were required for a minimal description. Speakers with such a strategy could hence begin describing earlier, before they had completely planned their utterance or even inspected the full visual scene (see Pechmann, 1989 and Gatt, Krahmer, Van Deemter, & Van Gompel, 2017 for similar results). In other words, in experimental setups like ours, where there is a well-controlled set of attributes that are manipulated, producing an overspecified description may actually be easier in terms of *message planning* (even though it is more difficult in terms of articulatory effort) than producing a minimal description. This will especially be the case if overspecification is part of a global strategy employed throughout the experiment. Thus, although overspecification may have facilitated production-internal processes on a trial-by-trial basis, this does not exclude the possibility that the speaker's decision to use an overspecification strategy in the first place is motivated by audience design.

Secondly, we did not find evidence that overspecified descriptions helped drivers to select the correct referent more quickly in the difficult driving task. In addition, driving accuracy did not improve when descriptions were overspecified, suggesting that overspecification did not benefit driving performance either. It could be the case that listening to a longer description impaired performance on the driving task. It is also possible that any facilitating effect was countered by an effect in the other direction: Whenever the driver was having trouble with the steering task, speakers might have responded to that by including more redundancy. Our results are therefore not conclusive as to whether overspecification is beneficial for the addressee or not. More detailed analyses of individual task strategies are needed to clarify this issue. For example, in future research we could analyse listeners' eye movements to investigate how their processing of the descriptions may be affected by increased redundancy (cf. Engelhardt et al., 2006; Tourtouri et al., 2017). However, even if increasing redundancy is not helpful for listeners, this does not preclude the possibility that the use of redundant descriptions is still *intended* by the speaker to accommodate the listener.

Although production-internal processes may have played a role in our study, they are not sufficient to explain our findings on the rate of overspecification. Most importantly, it is not clear why an increased use of redundant descriptions for speaker-internal reasons would only occur in speakers who had done the dual task themselves first and received the difficult driving condition as the first driving task. The order in which the two driving conditions were presented should not matter if speakers were only driven by their own preferences. At this point, therefore, we believe that the most likely explanation of our data is that speakers at the start of their description task chose a global strategy to overspecify their descriptions in order to aid their addressees, and only did so when it was evident to them that some adjustment was in order (i.e. when they had experienced the driver's cognitive load themselves). To further separate speaker-internal processes from audience design, future research could investigate speaker disfluencies and self-repairs in overspecifications. If speakers overspecify because they are having difficulty with the referential task themselves, one should find a higher rate of disfluency in these descriptions (cf. Arnold, 2008).

Our results are consistent with the findings of a comparable referential communication experiment using the same furniture stimuli as used in the current study by Koolen et al. (2011), who did not find significant effects of the communicative setting on the degree of overspecification. In one of their conditions, participants produced written descriptions without an addressee; in a second and third condition, participants produced spoken descriptions to a confederate addressee, but only in the third condition the speaker and addressee could also see each other. Koolen et al. predicted that speakers would produce more redundant attributes in spoken than in written descriptions, as well as when speaker and addressee cannot see each other vs. when they can, arguing that it is more important to be clear for the addressee when possibilities for feedback are limited. Although these predictions were borne out numerically in the data, the

differences between the conditions were not significant. It is possible that the cues for the speaker to consider the addressee's needs were not strong enough in that study. In addition, it is not obvious that introducing more redundancy will always increase clarity of the description for the listener, as overspecified descriptions may also be more confusing (e.g. Engelhardt et al., 2006; Sedivy, Tanenhaus, Chambers, & Carlson, 1999).

In our study, we used a stronger cue for adaptation than just co-presence of speaker and addressee, namely the cognitive status of the addressee. We expected that a noticeably cognitively loaded listener would be more likely to elicit audience design in the speaker to secure successful communication. In addition, we reasoned that an increase in the rate of overspecification leads to a decrease in information density, which may be beneficial for addressees with a reduced processing capacity. Still, our results suggest that even loading the addressee with a difficult dual task may not be sufficient to elicit audience design, but that the speaker apparently also needs to have first-hand evidence that the task is really impeding the capacity to process information. Moreover, a chosen strategy may not be easily abandoned in the absence of clear evidence that it is not working. Given the listener-driver's overall high response accuracy scores, speakers in our study might not have had a good enough reason to change their initial referential strategy.

To sum up, our findings do not seem to be in line with a strong interpretation of the Uniform Information Density Hypothesis (e.g. Levy & Jaeger, 2007) in which speakers distribute information uniformly over their utterances to ease the processing of that information by their addressees. UID provided us with a clear prediction about what audience design should look like if it was going to aid a cognitively loaded listener: A listener whose capacity for processing incoming linguistic material is clearly compromised by the performance of a secondary task will benefit from a lower rate of transmission of the linguistic information. Hence, introducing more redundancy that spreads out the crucial information over more time will be a good way for the speaker to aid the listener's linguistic processing. This is not what speakers did in our study: They did not adapt their degree of redundancy to a change in the addressee's cognitive load, nor did they more generally reduce the information density of their descriptions, by distributing information over more words or over more time.

However, our results also cannot be taken as evidence that speakers do not take into account the cognitive state of their addressees when choosing referring expressions. Firstly, the degree and type of listener adaptation is likely to be dependent on the specific task. For example, compared to the present study, a referential task in which finding the correct referent is much more difficult, or in which the importance of performing this task correctly is emphasised might elicit stronger adaptation effects. Secondly, to the extent that audience design is cognitively effortful, speakers may have more cognitive resources available to consider the addressee's needs when their own description task is easier. Thirdly, another factor that might play a role is the predictability of the referent (cf. Rosa et al., 2015). When the referent is relatively predictable, speakers may normally use an informationally dense description, and can then easily lower the information density when addressee needs require it. When the referent is not predictable to start with, speakers are likely to use a lower information density anyway, and addressee effects may be smaller. In our study, which object would be described was not predictable for the addressee, and hence we might have seen stronger accommodation effects if descriptions would have been more predictable. Finally, and most importantly, we found that the group of participants who had experienced the driving task themselves and who were presented with the difficult driving task first produced more redundant descriptions than other participants. This behaviour might constitute a referential strategy involving adjustment to the listener's needs at a more global level, and hence is not necessarily incompatible with rational pragmatic theories of communication such as UID or the Rational Speech Act model.

Notes

1. These stimuli have been used to collect the English TUNA (Gatt, Van Der Sluis, & Van Deemter, 2007) and Dutch D-TUNA (Koolen & Krahmer, 2010) corpora of referring expression production. One of the goals of the current research was to create a comparable corpus for German referring expression production. This corpus is described in Howcroft, Vogels, and Demberg (2017).
2. Stimulus images courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University, http://www.tarrlab.org/. This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.
3. No random slopes for the predictor Role order were included, since whether the speaker had described first or driven first was manipulated between-participants and between-items. In addition, random slopes for the control predictors were not included, as these are considered non-essential (Barr, Levy, Scheepers, and Tily, 2013). Adding them afterwards did not significantly improve model fit.
4. Alternatively, according to Schober (1993), speakers may only consider the listener's needs when the listener had

considered their needs when he was the speaker, as a kind of quid pro quo arrangement.

Acknowledgements

This work was supported by the German Research Foundation (DFG) as part of SFB 1102 "Information Density and Linguistic Encoding", and the Netherlands Organisation for Scientific Research (NWO), under Grant 275-89-0360. We are grateful to Fabio Lu, Matthias Lindemann, Ben Peters, Margarita Ryzhova and Katja Häuser for their help in collecting and analysing the data, to Stefan Ecker for software development and modification, to the OpenDS team for their assistance in modifying their system to suit our needs, and to Ruud Koolen for making the TUNA stimuli available for us.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the German Research Foundation (DFG) as part of SFB 1102 "Information Density and Linguistic Encoding", and the Netherlands Organisation for Scientific Research (NWO), under Grant 275-89-0360. We are grateful to Fabio Lu, Matthias Lindemann, Ben Peters, Margarita Ryzhova and Katja Häuser for their help in collecting and analysing the data, to Stefan Ecker for software development and modification, to the OpenDS team for their assistance in modifying their system to suit our needs, and to Ruud Koolen for making the TUNA stimuli available for us.

ORCID

Jorrig Vogels http://orcid.org/0000-0001-6698-504X
David M. Howcroft http://orcid.org/0000-0002-0810-9065
Elli Tourtouri http://orcid.org/0000-0003-0453-7377
Vera Demberg http://orcid.org/0000-0002-8834-0020

References

Arnold, J. E. (2008). Reference production: Production-internal and addressee-oriented processes. *Language and Cognitive Processes*, *23*(4), 495–527. doi:10.1080/01690960801920099

Arnold, J. E., & Griffin, Z. M. (2007). The effect of additional characters on choice of referring expression: Everyone counts. *Journal of Memory and Language*, *56*(4), 521–536. doi:10.1016/j.jml.2006.09.007

Arnold, J. E., Kahn, J. M., & Pancani, G. C. (2012). Audience design affects acoustic reduction via production facilitation. *Psychonomic Bulletin & Review*, *19*(3), 505–512. doi:10.3758/s13423-012-0233-y

Arts, A., Maes, A., Noordman, L. G. M., & Jansen, C. (2011). Overspecification in written instruction. *Linguistics*, *49*(3), 555–574. doi:10.1515/ling.2011.017

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language*, *42*(1), 1–22. doi:10.1006/jmla.1999.2667

Bard, E. G., & Aylett, M. P. (2005). Referential form, word duration, and modeling the listener in spoken dialogue. In J. Trueswell & M. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions* (pp. 173–191). Cambridge, MA: MIT Press.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. doi:10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *ArXiv Preprint ArXiv:1506.04967*.

Becic, E., Dell, G. S., Bock, K., Garnsey, S. M., Kubose, T., & Kramer, A. F. (2010). Driving impairs talking. *Psychonomic Bulletin & Review*, *17*(1), 15–21. doi:10.3758/PBR.17.1.15

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482. doi:10.1037/0278-7393.22.6.1482

Brennan, S. E., & Hanna, J. E. (2009). Partner-specific adaptation in dialog. *Topics in Cognitive Science*, *1*(2), 274–291. doi:10.1111/j.1756-8765.2009.01019.x

Christensen, R. H. B. (2015). *A Tutorial on fitting Cumulative Link Mixed Models with clmm2 from the ordinal Package*. The Comprehensive R Archive Network.

Clark, H. H. (1996). *Using language*. Cambridge: University Press.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. doi:10.1016/0010-0277(86)90010-7

Dell, G. S., & Brown, P. M. (1991). Mechanisms for listener-adaptation in language production: Limiting the role of the "model of the listener". In D. Napoli, & J. A. Kegl (Eds.), *Bridges between psychology and linguistics: A Swarthmore Festschrift for Lila Gleitman* (pp. 105–129). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210. doi:10.1016/j.cognition.2008.07.008

Demberg, V., Sayeed, A., Mahr, A., & Müller, C. (2013). Measuring linguistically-induced cognitive load during driving using the ConTRe task. In *Proceedings of the 5th international conference on automotive user interfaces and interactive vehicular applications* (pp. 176–183). New York, NY, USA: ACM. doi:10.1145/2516540.2516546

Drews, F. A., Pasupathi, M., & Strayer, D. L. (2008). Passenger and cell phone conversations in simulated driving. *Journal of Experimental Psychology: Applied*, *14*(4), 392–400. doi:10.1037/a0013119

Engelhardt, P. E., Bailey, K. G. D., & Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, *54*(4), 554–573. doi:10.1016/j.jml.2005.12.009

Engelhardt, P. E., & Ferreira, F. (2014). Do speakers articulate over-described modifiers differently from modifiers that are required by context? Implications for models of reference production. *Language, Cognition and Neuroscience*, *29*(8), 975–985. doi:10.1080/01690965.2013.853816

Engonopoulos, N., Sayeed, A., & Demberg, V. (2013). Language and cognitive load in a dual task environment.

In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2249–2254). Austin, TX: Cognitive Science Society.

Federmeier, K. D., McLennan, D. B., De Ochoa, E., & Kutas, M. (2002). The impact of semantic memory organization and sentence context information on spoken language processing by younger and older adults: An ERP study. *Psychophysiology*, *39*(2), 133–146. doi:10.1017.S004857720139203X

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language Games. *Science*, *336*(6084), 998–998. doi:10.1126/science.1218633

Fukumura, K., & van Gompel, R. P. (2012). Producing pronouns and definite noun phrases: Do speakers use the addressee's discourse model? *Cognitive Science*, *36*(7), 1289–1311. doi:10.1111/j.1551-6709.2012.01255.x

Galati, A., & Brennan, S. E. (2010). Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language*, *62*(1), 35–51. doi:10.1016/j.jml.2009.09.002

Gann, T. M., & Barr, D. J. (2014). Speaking from experience: Audience design as expert performance. *Language, Cognition and Neuroscience*, *29*(6), 744–760. doi:10.1080/01690965.2011.641388

Gatt, A., Krahmer, E., Van Deemter, K., & van Gompel, R. P. (2017). Reference production as search: The impact of domain size on the production of distinguishing descriptions. *Cognitive Science*, *41*, 1457–1492. doi:10.1111/cogs.12375

Gatt, A., Van Der Sluis, I., & Van Deemter, K. (2007). Evaluating algorithms for the generation of referring expressions using a balanced corpus. *Proceedings of the Eleventh European Workshop on Natural Language Generation*, 49–56. Association for Computational Linguistics.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. doi:10.1016/j.tics.2016.08.005

Goudbeek, M., & Krahmer, E. (2011). Referring under load: Disentangling preference-based and alignment-based content selection processes in referring expression generation. In K. van Deemter, A. Gatt, R. P. G. van Gompel, & E. Krahmer (Eds.), *Proceedings of PRE-Cogsci: Bridging the gap between computational, empirical and theoretical approaches to reference*. Boston, MA: Cognitive Science Society.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics. Vol III: Speech acts* (pp. 41–58). New York: Academic Press.

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, *69*, 274–307. doi:10.2307/416535

Hendriks, P. (2016). Cognitive modeling of individual variation in reference production and comprehension. *Frontiers in Psychology*, *7*, 506. doi:10.3389/fpsyg.2016.00506

Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition*, *96*(2), 127–142. doi:10.1016/j.cognition.2004.07.001

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*(1), 91–117. doi:10.1016/0010-0277(96)81418-1

Howcroft, D., Vogels, J., & Demberg, V. (2017). G-TUNA: A corpus of referring expressions in German, including duration information. *Proceedings of the 10th International Conference on Natural Language Generation*, 149–153. doi:10.18653/v1/W17-3522

Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General*, *116*(1), 26–37.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62. doi:10.1016/j.cogpsych.2010.02.002

Jaeger, T. F., & Tily, H. (2011). On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(3), 323–335. doi:10.1002/wcs.126

Jucks, R., Becker, B.-M., & Bromme, R. (2008). Lexical entrainment in written discourse: Is experts' word Use adapted to the addressee? *Discourse Processes*, *45*(6), 497–518. doi:10.1080/01638530802356547

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, *43*(13), 3231–3250. doi:10.1016/j.pragma.2011.06.008

Koolen, R., Goudbeek, M., & Krahmer, E. (2013). The effect of scene variation on the redundant use of color in definite reference. *Cognitive Science*, *37*(2), 395–411. doi:10.1111/cogs.12019

Koolen, R., & Krahmer, E. (2010). The D-TUNA Corpus: A Dutch dataset for the evaluation of referring expression generation algorithms. *LREC*. Retrieved from http://tst-centrale.org/images/stories/producten/documentatie/dtuna_documentatie_en.pdf

Krauss, R. M., & Fussell, S. R. (1991). Perspective-taking in communication: Representations of others' knowledge in reference. *Social Cognition*, *9*(1), 2–24. doi:10.1521/soco.1991.9.1.2

Kuhlen, A. K., & Brennan, S. E. (2010). Anticipating distracted addressees: How speakers' expectations and addressees' feedback influence storytelling. *Discourse Processes*, *47*(7), 567–587. doi:10.1080/01638530903441339

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In M. Bar (Ed.), *Predictions in the brain: Using our past to generate a future* (pp. 190–207). Oxford, New York: Oxford University Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. doi:10.1016/j.cognition.2007.05.006

Levy, R. P., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. C. Platt, & T Hoffman (Eds.), *Advances in neural information processing systems* (pp. 849–856).

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318. doi:10.1016/j.cognition.2012.09.010

Mahr, A., Feld, M., Moniri, M. M., & Math, R. (2012). The contre (continuous tracking and reaction) task: A flexible approach for assessing driver cognitive workload with high sensitivity. *Automotive User Interfaces and Interactive Vehicular Applications*, 88–91.

Math, R., Mahr, A., Moniri, M. M., & Müller, C. (2012). OpenDS: A new open-source driving simulator for research. *Proceedings of the International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Adjunct Proceedings*, 7–8.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*(1), 89–110. doi:10.1515/ling.1989.27.1.89

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291. doi:10.1016/j.cognition.2011.10.004

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190. doi:10.1017/S0140525X04000056

Rosa, E. C., Finch, K. H., Bergeson, M., & Arnold, J. E. (2015). The effects of addressee attention on prosodic prominence. *Language, Cognition and Neuroscience*, *30*(1–2), 48–56. doi:10.1016/j.jml.2016.07.007

Rossnagel, C. (2000). Cognitive load and perspective-taking: Applying the automatic-controlled distinction to verbal communication. *European Journal of Social Psychology*, *30*(3), 429–445. doi:10.1002/(SICI)1099-0992(200005/06)30:3≤429::AID-EJSP3≥3.0.CO;2-V

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition*, *47*, 1–24. doi:10.1016/0010-0277(93)90060-9

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147. doi:10.1016/S0010-0277(99)00025-6

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423, 623–656. doi:10.1002/j.1538-7305.1948.tb01338.x

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. doi:10.1016/j.cognition.2013.02.013

Tourtouri, E. N., Delogu, F., & Crocker, M. W. (2017). Specificity and entropy reduction in situated referential processing. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th annual conference of the cognitive science society* (pp. 3356–3361). Austin, TX: Cognitive Science Society. Retrieved from https://pdfs.semanticscholar.org/6b3e/88f697b2850fa9019ea17f15c2675cb306f8.pdf

Van Berkum, J. J. (2008). Understanding sentences in context: What brain waves can tell us. *Current Directions in Psychological Science*, *17*(6), 376–380. doi:10.1111/j.1467-8721.2008.00609.x

Vogels, J., Krahmer, E., & Maes, A. (2015). How cognitive load influences speakers' choice of referring expressions. *Cognitive Science*, *39*(6), 1396–1418. doi:10.1111/cogs.12205

Wardlow Lane, L., Groisman, M., & Ferreira, V. S. (2006). Don't talk about pink elephants!: speakers' control over leaking private information during language production. *Psychological Science*, *17*(4), 273–277. doi:10.1111/j.1467-9280.2006.01697.x

Watson, D. G., Arnold, J. E., & Tanenhaus, M. K. (2008). Tic Tac TOE: Effects of predictability and importance on acoustic prominence in language production. *Cognition*, *106*(3), 1548–1557. doi:10.1016/j.cognition.2007.06.009

Xiang, M., & Kuperberg, G. (2015). Reversing expectations during discourse comprehension. *Language, Cognition and Neuroscience*, *30*(6), 648–672. doi:10.1080/23273798.2014.995679

Zipf, G. K. (1949). *Human behaviour and the principle of least-effort*. Reading: Addison-Wesley.